Primer: Understanding Systematic Errors in Phylogenetic Trees.

Paschalia Kapli[1], Tomáš Flouri[1], Maximilian J Telford[1]*

1. Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

* email: m.telford@ucl.ac.uk

The effort to reconstruct the tree of life was revolutionized by the development of methods to determine the sequence of proteins and nucleic acids. Phylogenetic trees are now routinely reconstructed using hundreds of thousands of amino acid or nucleotide characters and it seems surprising, therefore, that many aspects of the tree of life are still controversial but conflicting results between large scale phylogenomic studies show that errors in the tree of life remain common. These errors often result from unaccounted for systematic biases in the way sequences evolve and, while the resulting systematic errors are well understood, it requires careful efforts to reduce their effects.

The fundamentals of molecular phylogenetics are straightforward: by aligning homologous genes (those inherited from a common ancestor), individual homologous nucleotides and amino acids can be identified in different species. The heritable substitutions that these sites in a gene/protein experience in different lineages over evolutionary history are passed on to their descendants and constitute a record of species relationships.

The most common source of error in phylogenetic reconstruction is homoplasy, whereby the same novel character appears in two species through convergent evolution rather than because both have inherited it from their common ancestor. Homoplasy is inevitable in molecular sequence data due to the limited set of character states any given site in an alignment can adopt – 4 different nucleotides and 20 different amino acids. Homoplasy results in two types of error: stochastic error resulting from small samples (and mostly eliminated by large data sets); and the more devious systematic error.

Systematic error is a consistent and repeatable error that results from faulty assumptions in our analysis. Such errors commonly arise when the models we use to reconstruct our trees assume that the process of sequence change is homogeneous when in reality the process is heterogeneous - either across sites, between taxa, or through time. Ignoring these heterogeneities sometimes results in artifacts in phylogenetic inference and we are more certain to see an error caused by systematic error as we add more data.

We will focus on understanding three of the best-known types of heterogeneity in sequence evolution which, when ignored, may result in phylogenetic inference errors: i) heterogeneous rates of evolution across lineages, ii) heterogeneous rate of evolution across sites within an alignment, and iii) heterogeneous state composition across sites. We use three simple and short synthetic sequence alignments to illustrate how these ubiquitous features of molecular evolution can cause errors in reconstructing trees.

Heterogeneous rates across taxa

Evolutionary rates (the probability that a nucleotide or amino acid will change from one state to another in a given time period) can vary substantially between species, especially between more distantly related taxa. Two unrelated, fast-evolving organisms will each tend to accumulate many mutations and, given the limited diversity of nucleotides/amino acids, there is a risk that both will evolve the same new state at a given site of a gene/protein. Their slow evolving relatives are less likely to experience this convergence. If this asymmetry in outcomes is not accounted for, there is the risk of misinterpreting homoplasy for true phylogenetic signal and

clustering unrelated species together. The phenomenon whereby unrelated fast evolving taxa are incorrectly grouped is infamous in phylogenetics as the Long Branch Attraction (LBA) artefact.

In our examples, we assume just such a phylogeny involving two relatively slowly evolving species - an arthropod and a vertebrate - and two rapidly evolving species - a nematode and a ctenophore (comb jelly) which is the outgroup to the other three species. The correct tree groups the fast-evolving nematode with the slowly evolving arthropod. The nematode and arthropod are separated from the vertebrate and the outgroup by a short branch that contains little evidence of the separation. We show that when the variation in the expectation of change across taxa or sites is not accounted for, we obtain the LBA topology whereby the long-branched nematode is attracted to the long branch leading to the outgroup. The corollary of this is that the short-branched arthropod is incorrectly placed as the closest relative of the short-branched vertebrate.

Maximum Parsimony trees and an illustration of LBA

The maximum parsimony (MP) tree building method treats all sites in an alignment and all taxa equally and therefore does not explicitly model the potential difference in evolutionary rates between taxa. MP chooses the tree topology that minimises the number of substitutions required to explain the observed data. Only certain site patterns imply different numbers of changes on different tree topologies and any characters that are not 'parsimony informative' in this way are ignored by the method. In our example of seven homologous sites (Fig. 1A), only the first three sites are parsimony informative; in the last four, the changes have occurred along the branches leading to a single taxon (importantly for our purposes, these changes have occurred in the two fast evolving taxa). These last site patterns can be explained by a single substitution on any topology meaning they are uninformative.

Of the parsimony informative characters, one (indicated by a red triangle) supports the correct topology as the arthropod and the nematode share the same nucleotide, inherited from a mutation that occurred in their common ancestor. In contrast, in the next two sites (indicated by blue stars) the two fast-evolving branches (nematode and ctenophore) have the same nucleotide which evolved through independent mutations in the two branches. (Fig 1A).

When we compare the number of changes in the three parsimony informative characters on two possible tree topologies, we see that the true tree requires 5 changes to explain the observed sequence data (Fig 1B). The two convergently evolved characters (blue) must each change twice plus a single change in the red character. On the LBA tree, in contrast, the two blue characters are both interpreted as having changed once on the branch separating the ctenophores and nematodes from the vertebrate and arthropod. Only the red character is interpreted as having evolved convergently (and is therefore counted as experiencing two changes). The LBA tree requires 4 changes to explain the observed character distribution and is therefore the preferred maximum parsimonious tree.

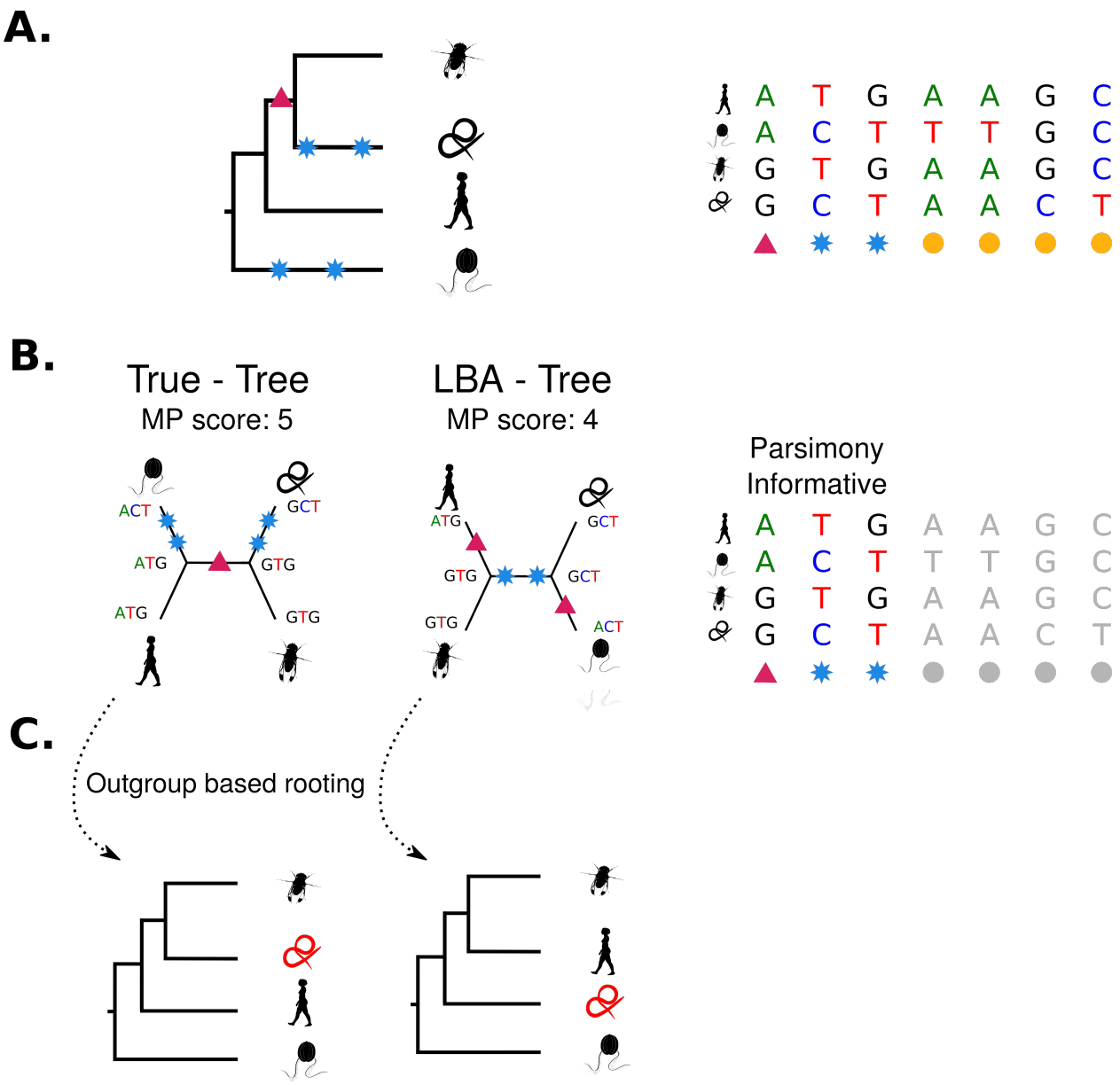

Figure 1. Parsimony and long branch attraction.

A. The tree on the left shows the true topology underlying all our examples. The ctenophore is the outgroup and serves to root the subtree containing the three bilaterian animals. Within the Bilateria, the vertebrate is the sister group of the Ecdysozoan clade consisting of nematodes and arthropods. On the right is an alignment of homologous nucleotides that have evolved along this phylogeny; it consists of seven variable sites of which only the first three are parsimony informative. In the first site, the insect and the nematode share a character state and the vertebrate and ctenophore share another - this arose as a single change (red triangle on tree). In the second and third sites, the chordate and the insect share one nucleotide while the nematode and the ctenophore share another. This pattern has arisen through convergent evolution of the same character state in the two divergent taxa - nematode and ctenophore (blue stars). In each of the four remaining sites (denoted by orange dots below the alignment) a single taxon has a unique nucleotide (changes that occurred in the fast-evolving taxa) and are ignored by the MP method.

B. We consider the minimum number of changes required to explain these site patterns on two alternative topologies (we show unrooted trees). The tree on the left gets a parsimony score of 5. This correctly accounts for the single change that occurred in the red character and the two changes in each of the two blue characters. The topology on the right gets a lower parsimony score (and is therefore preferred) as it assumes convergence of the red character (independent changes in arthropod and vertebrate) but a single change in each of the blue characters in the branch separating nematode and ctenophore from arthropod and vertebrate. This tree clusters the two fast evolving taxa (nematode and ctenophore) together and is considered to be the LBA tree

C. Maximum parsimony and other phylogenetic inference methods infer unrooted phylogenies as shown in B. In the presence of an outgroup (the ctenophore here) we can root the phylogeny. The rooted phylogeny on the left corresponds to the true tree while the one on the right differs in placing the nematode closer to the root.

Likelihood methods are branch length aware and overcome LBA.

The likelihood-based methods (Maximum Likelihood and Bayesian approaches) use an explicit model of sequence evolution to estimate the probability of observing the sequences in the species on a given tree topology with a certain set of branch lengths. Branch lengths reflect the expected number of changes per site in a given dataset and a branch may be long because a lot of time has passed or because the rate of change is fast. To give a simple example, two species with very similar sequences will have a high likelihood of being separated by a short branch and a low likelihood of being separated by a long branch. Figure 2A shows that there is greater likelihood for the arthropod to be separated from its ancestor by a short branch reflecting the few changes it has accumulated. On the contrary, for the fast-evolving nematode, a longer branch has a higher likelihood.

Knowing that a branch is long from the totality of data also feeds back on our expectations of change for individual sites. In the plot shown in Figure 2B we can see that the probability of a given nucleotide changing to any other nucleotide increases with increasing branch length. If a species is at the end of a long branch this means that any one of its characters is more likely to have changed along that branch than if it is at the end of a short branch.

The effect on tree reconstruction from accounting for branch lengths is shown in Figure 2C. Using the same data as in Figure 1, we see that the difference in the likelihood of the two tree topologies changes as we include more of the parsimony non informative characters. Under likelihood, these parsimony uninformative characters provide information about the long branches leading to the nematode and the ctenophore. As we add more of the characters which have changed in the long branch taxa, these branches get longer. The likelihood of a change along the two long branches increases and so a topology in which the homoplastic characters have changed twice becomes more likely. As we add more sites that reveal the high rate of change in the long branches, we see that the likelihood of the true tree becomes higher than the likelihood of the LBA tree. Parsimony effectively underestimates the likelihood of change in sites along long branches and hence underestimates the likelihood of convergence. Branch length aware likelihood models can accommodate branch length heterogeneities and hence reconstruct the correct topology.

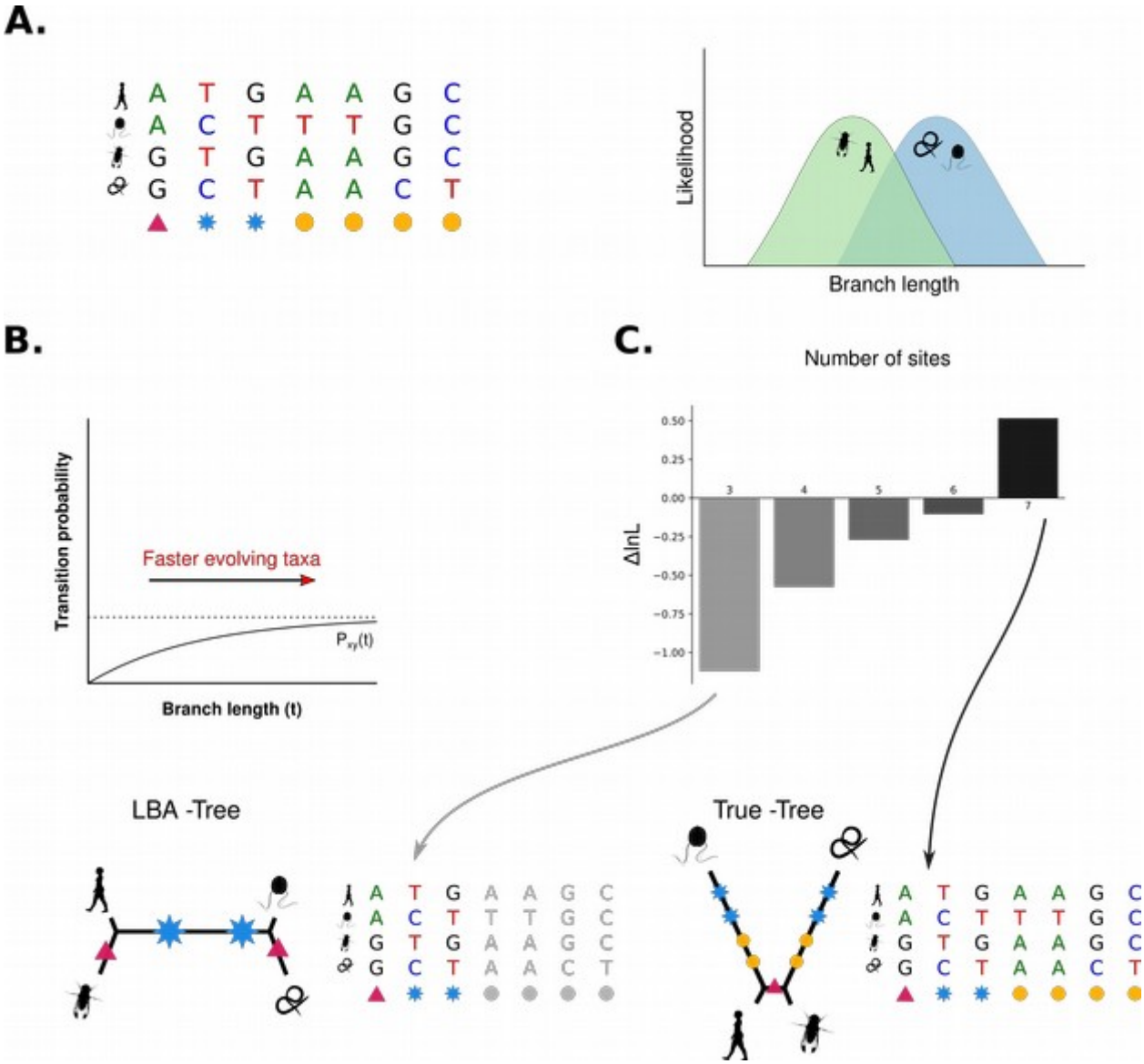

Figure 2. Likelihood methods accommodate branch length variance and avoid LBA artefacts.

A. On the left we see the same sequence alignment as in Figure 1. On the right is a cartoon of a graph representing the likelihoods (y axis) of observing the data in the alignment given different lengths (x axis) of the branches leading to arthropods/vertebrates (green) and nematodes/ctenophores (blue). The observed data are most likely to have been produced if the sequences evolved along a tree in which the arthropod and the vertebrate branches are short, and the nematode and the ctenophore branches are long.
B. Graph showing the change in probability of a transition from one nucleotide to another with increasing branch length (under the simple Jukes-Cantor model - equal frequencies of the 4 nucleotides and equal transition probability from each nucleotide to any other). The longer the branch, the more likely it is that the nucleotide has changed. Branches can be long because a lot of time has passed or because the rate of change per unit time is high.
C. The likelihoods of the two tree topologies change as we add sites that provide evidence that the branches leading to nematodes and ctenophores are long. The graph shows the difference in lnLikelihoods between the LBA tree (left) and the true tree (right) as we add more of the parsimony uninformative sites (orange). These contain information showing that the ctenophore and nematode have many changes/long branches. As more of the parsimony uninformative sites are added, the true tree becomes more likely (values > 0 support the true tree). As the branches become longer and the probability of any site experiencing a substitution increases, it becomes more plausible to infer independent changes in both nematodes and ctenophores. Likelihood methods are able to accommodate the between-species rate heterogeneity observed in these data and infer the correct tree.

Heterogeneities in rates between sites can be accommodated by likelihood models.

Apart from heterogeneities between species in the process of evolution, one can also observe heterogeneities between sites within an alignment. Perhaps the most obvious such heterogeneity is the difference in the rates of change between sites within a gene (or between genes within a concatenated alignment). Individual amino acids, protein domains and whole genes are under very different selective pressures and so will have changed at different rates. At the level of nucleotides in a coding sequence, changes to the third position of a codon are typically silent (do not result in a change in amino acids) and hence occur much more frequently than changes in the first two codon positions.

Combined with species rate heterogeneity (unequal branch lengths), variance in the rate of change between sites has been shown to cause likelihood models to fail to recover the correct topology. The failure to model differences in rates between sites - effectively having a single (average) rate for all sites - means that one systematically underestimates the likelihood of change in the faster-evolving sites and overestimates the likelihood of change in the slower evolving sites. Because the faster-evolving sites are those most likely to undergo convergent evolution, assuming a homogeneous rate of change across sites can result in LBA artefacts.

This problem is illustrated in Figure 3 where we consider a protein coding alignment in which the 3rd codon positions evolve more quickly than the 1st and 2nd positions. The homoplastic sites are most likely to occur in the fast evolving 3rd positions. If we ignore the rate variance and assume a uniform rate across all sites, we systematically underestimate the likelihood that the fast-evolving 3rd codon sites change and hence underestimate the likelihood that they might evolve convergently. In an alignment for which we model a single rate for all sites, we see that the likelihood method prefers the LBA tree. If, however, we partition the data and use one estimate of rate for 1st and 2nd positions and a second estimate of rate for the 3rd position, we find that, as we assume faster rates for position 3 relative to positions 1+2, the likelihood of the LBA tree decreases and that of the true tree increases. We can find the maximum likelihood estimate of the relative rates for the two partitions (the relative rates that have the highest likelihood of having produced the observed data) and when we use these maximum likelihood rates, the true tree has a higher likelihood than the LBA tree.

As a general method to accommodate this among-site rate variation, Yang (1993)-proposed the modelling of rates of sites as a random variable following a gamma distribution. This strategy for accounting for rate heterogeneity among sites is widely used.

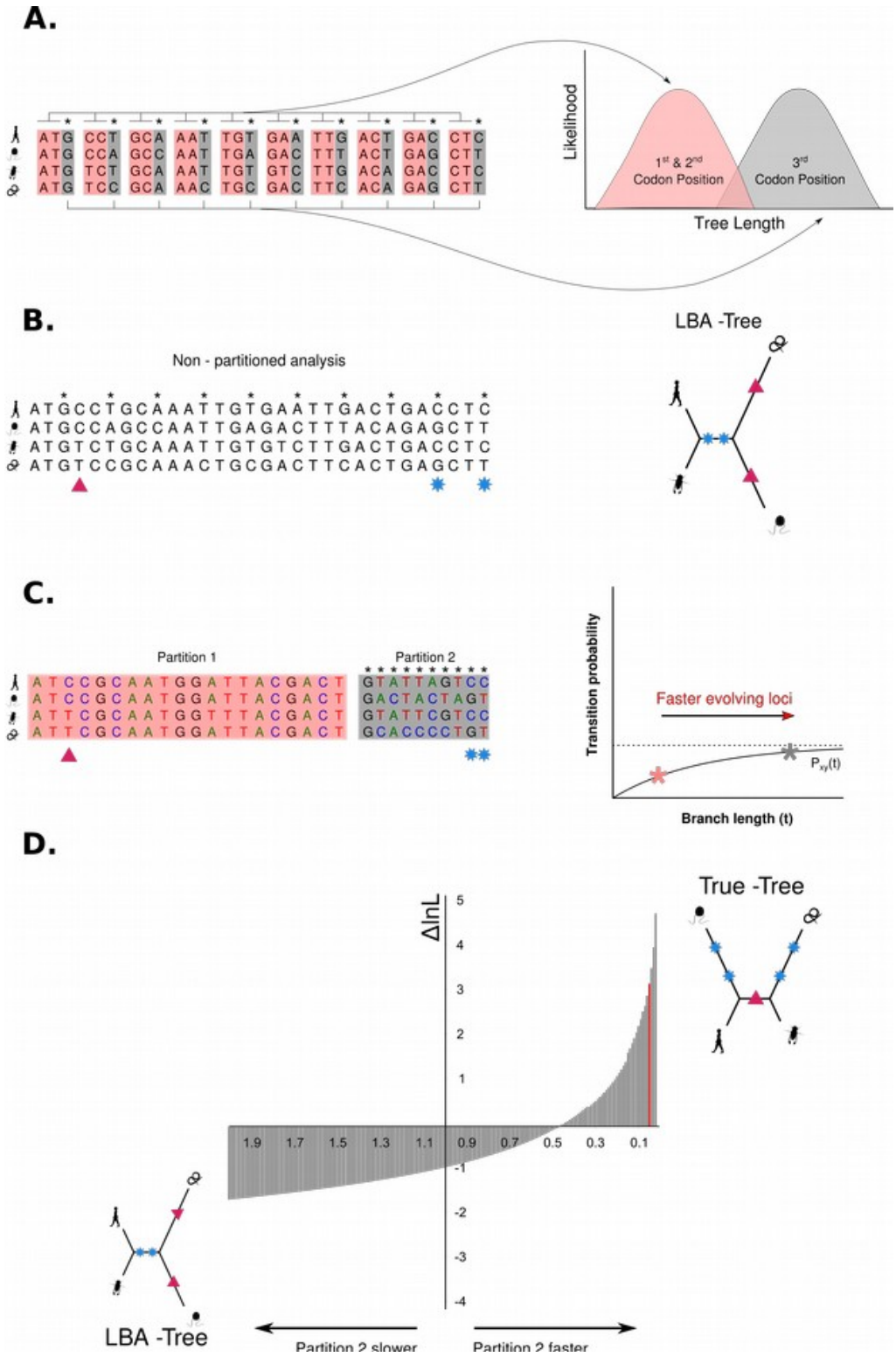

Figure 3. Modelling heterogeneities in rates between sites can help avoid LBA artefacts.

A. On the left is an alignment of a synthetic coding sequence from our four taxa; the alignment is separated into codons. 1st and 2nd position nucleotides (slow evolving) are shown in pink and 3rd position nucleotides (faster evolving) are shown in grey and annotated with an asterisk (similarly in B and C). On the right is a cartoon showing the likelihoods of observing the slower and the faster evolving partitions on trees of different lengths (slower or faster rates). The set of character states observed in the rarely changing 1st and 2nd position data are most likely to have been produced by sequences evolving along a shorter/slower tree; the faster changing 3rd position data are most likely to have been produced if the sequences evolved along a longer/faster tree. In this case, of course, the differences in tree length come from differences in rate and not in time elapsed.
B. Using unpartitioned data (a uniform, average rate assumed for all positions) the maximum likelihood tree is the LBA tree. Two 3rd codon position sites have undergone convergent evolution, but the use of a uniform rate

means the likelihood that these convergent changes occurred is underestimated. This is analogous to underestimating the likelihood of change (and hence of convergence) in more divergent taxa on the phylogeny.

C. Partitioning the data into slower and faster sites and separately estimating their rates of change can accommodate the site rate heterogeneity. The synapomorphy (red triangle) is in the slow partition and the homoplasies (blue stars) are in the fast partition. The graph on the right shows the change in probability of a transition from one nucleotide to another as we move from slower to faster rates of change (i.e., increasing tree length). The pink and the grey stars show the likelihood of change expected for the slow evolving 1st and 2nd position partition and the fast evolving 3rd position.

D. Graph showing the difference in the likelihoods (Delta lnL) of the true tree and the LBA tree when we assume different relative rates for the two partitions (x axis shows the ratios of the partition 1 rate to the rate of partition 2). When the rates of the two partitions are equal (x=1) the LBA tree is preferred (Delta lnL <0). Moving to the right on the graph we assume a higher rate of the 3rd position partition relative to the 1st/2nd position and the likelihood of the true tree increases (Delta lnL >1). When the rate of the 3rd position partition is more than twice the rate of 1st and 2nd position partition, the true topology is favoured. The maximum likelihood estimate of the relative rates is indicated by the red line (The true ratio of rates = 25:1) and at this point the true tree is the maximum likelihood tree.

Heterogeneities in composition between sites can be accommodated by likelihood models.

Different residues within a gene/protein not only change at different rates but are also frequently restricted to different subsets of all possible nucleotides or amino acids according to the function of that residue. A clear example would be two regions of a protein, one of which crosses a membrane and one of which is extracellular; the former is likely to be constrained to experience substitutions between different hydrophobic amino acids and the latter to be comprised predominantly of hydrophilic amino acids. Most models of sequence evolution, however, assume a uniform composition across residues within an alignment - they assume compositional homogeneity across sites. A model that uses the average rate of substitution between Leucine and Isoleucine residues measured across all sites of a data set, will tend to underestimate the real rate of substitution between these two amino acids at a position in the gene that, for functional reasons, can only accommodate a Leucine or Isoleucine.

We show the effects on tree reconstruction of ignoring this between site compositional heterogeneity in Figure 4. For simplicity, we consider a synthetic DNA sequence although compositional heterogeneity is probably more important for proteins. In our example, over the whole alignment the four nucleotides, ACGT, are found at equal frequencies, we have shown an extreme example of compositional heterogeneity, however, in that there are three categories of sites in the data. The first partition has equal frequencies of all four nucleotides; the second and third partitions, however, are constrained to contain either only GC nucleotides or only AT nucleotides respectively. We can see that the estimated probability of a transition between the states G and C is lower along a given branch length if all nucleotides are at equal frequency (0.25 each) than if the only character states possible are G or C (0.5 each).

In our example, assuming a uniform composition for the three partitions results in underestimating the likelihood of change in the second (GC rich) and third (AT rich) partitions. This leads to an underestimate of the probability of convergent evolution and as a result, the maximum likelihood tree is the LBA tree. When we partition the data according to their composition, we are able to obtain a correct estimate of the probability of transition for nucleotides in each partition and hence correctly identify the instances of convergent evolution on the two long branches. Under these conditions, the inferred maximum likelihood tree corresponds to the true tree.

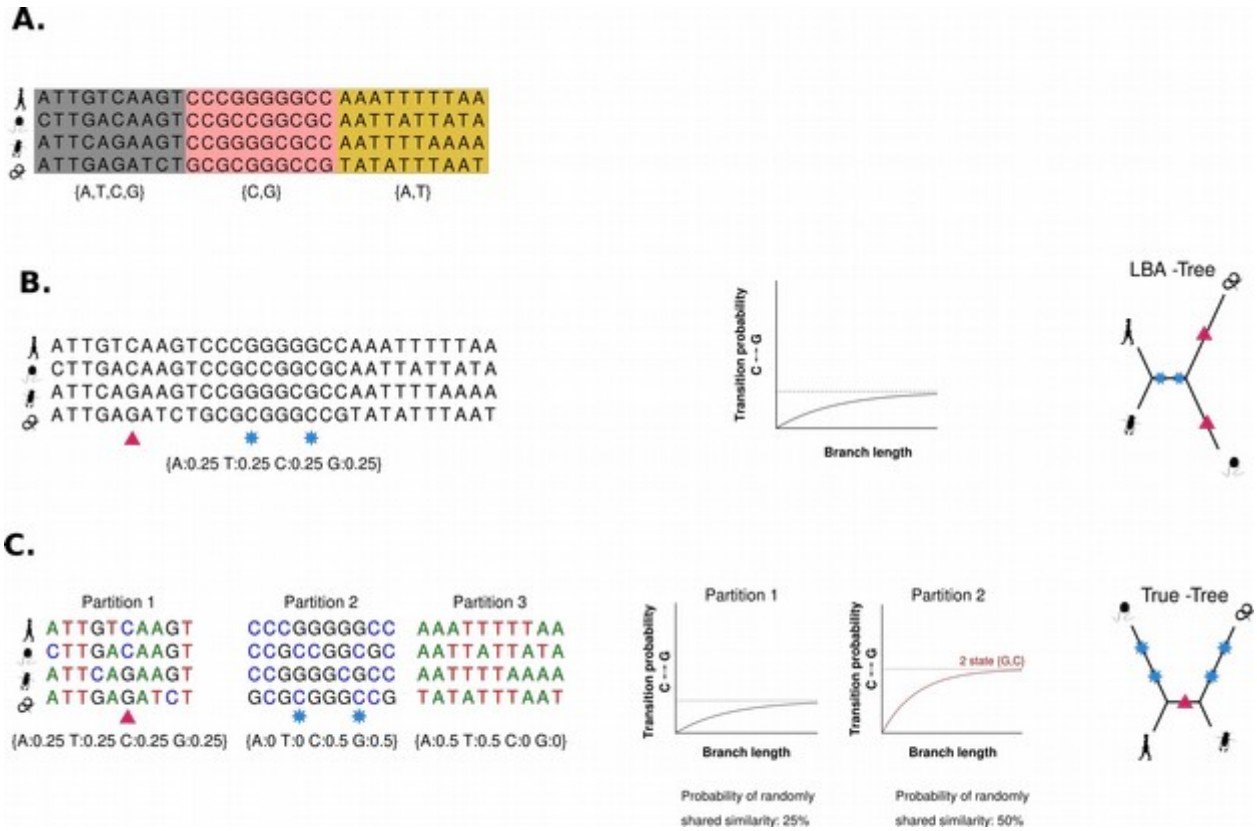

Figure 4. Modelling heterogeneities in composition between sites can help avoid LBA artefacts.

A. A synthetic alignment is shown that has three partitions, each with a different nucleotide composition. Over all three partitions, the four nucleotides have equal frequencies. The first partition also has equal composition of all four nucleotides. In the second partition there is a constraint that means that only G or C nucleotides are observed. In the third partition only A or T nucleotides are observed.

B. As in our other examples, there are three parsimony informative characters. The non-homoplastic red change is in partition 1 and the two homoplastic green substitutions are in partition 2, which is constrained to G or C nucleotides. The graph shows the probability of a G to C transition with increasing branch lengths under the assumption that all four nucleotides have an equal frequency (compositional homogeneity across sites as measured across the whole alignment). Because we underestimate the true probability of a G to C transition in partition 2, we underestimate the likelihood that the changes in the green sites are homoplastic. Under the assumption of compositional homogeneity across the alignment, the maximum likelihood tree is the LBA topology.

C. We can partition the data into the three partitions and estimate the frequencies of each nucleotide separately for each partition. The graphs show the probability of a G to C transition with increasing branch lengths under the assumptions i) that all four nucleotides have an equal frequency (partition 1) and ii) that G and C have a frequency of 0.5 each (partition 2). The same is true for A and T in partition 3. When we partition the data and assume the true frequencies of G and C, the higher probability of a G to C transition in partition 2 is correctly estimated and hence the likelihood of homoplasy is correctly calculated. When we partition the data to account for the compositional heterogeneity amongst partitions, the maximum likelihood tree is the true tree.

Our simple examples are designed to show how ignoring rate heterogeneities and compositional variance may result in errors in phylogenetic inference. We have shown that partitioning the data into homogeneous subsets and applying suitable models to each of them can overcome these errors. For real data, however, we don't necessarily know *a priori* how to partition data and we rely on mixture models which assume multiple sets of parameters for each site and average over their likelihoods. Among these models are the Gamma model which accommodates across site rate variation and the CAT (categories) and C10-C60 models, which accommodate across site compositional variance. These models have been shown to make correct estimates of branch lengths for heterogeneously evolving data and also to suppress the long-branch attraction artefacts that might result.

While we have described three important sources of systematic error, empirical studies have revealed several other types of deviations from homogeneous sequence evolution that can result in phylogenetic inference errors. Composition can vary across taxa (e.g. having an AT or GC rich genome) and the rate of substitution and the composition of a given site may change across time (heterotachy and heteropecilly). These heterogeneities are harder to model and currently the most common strategy is to try to identify and remove data suffering from such biases.

The choice of data and models becomes increasingly important when inferring phylogenies that involve distantly related species and there are heated debates concerning various nodes of the tree of life that revolve around the suspicion of systematic error. Famous examples include whether the amitochondriate microsporidia are members of the fungi; the monophyly of the Ecdysozoa including nematodes and arthropods) and the affinities of the Ctenophora (sea gooseberries/comb jellies) and the xenacoelomorph worms. The models required to accommodate the heterogeneities of real sequence data in such cases are complex and the computational resources required to solve them therefore considerable. Without the effort to anticipate and to accommodate systematic errors, however, we cannot resolve the most difficult parts remaining on the tree of life.